

# 机器学习数学笔记

宋佳欢

2019 年 8 月 11 日

## 目录

<b>1 线性代数</b>	<b>1</b>
1.1 向量	1
1.2 矩阵	2
1.3 特征值与特征向量	3
1.4 矩阵和向量的求导	3
<b>2 概率论</b>	<b>4</b>
2.1 条件概率	4
2.2 贝叶斯公式	4
2.3 随机事件的独立性	4
2.4 期望与方差	5
2.5 常用分布	5
2.6 随机向量	5
2.7 协方差	5
2.8 最大似然估计	6
<b>3 优化问题</b>	<b>6</b>
3.1 拉格朗日乘子法	6

## 1 线性代数

### 1.1 向量

(1) 向量相加减:

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \pm \begin{pmatrix} d \\ e \\ f \end{pmatrix} = \begin{pmatrix} a \pm d \\ b \pm e \\ c \pm f \end{pmatrix}$$

(2) 向量内积:

$$(a \ b \ c) \begin{pmatrix} d \\ e \\ f \end{pmatrix} = a \times d + b \times e + c \times f$$

(3) 向量的范数:

$$\|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

L1 范数:

$$\|x\|_1 = \sum_{i=1}^n |x_i|$$

L2 范数 (向量的模):

$$\|x\|_2 = \sqrt{\sum_{i=1}^n (x_i)^2}$$

## 1.2 矩阵

(1) 矩阵乘法:

$$\begin{cases} (AB)C = A(BC) \\ (A+B)C = AC + BC \\ A(B+C) = AB + AC \\ (AB)^T = B^T A^T \end{cases}$$

(2) 逆矩阵:

$$\begin{cases} (AB)^{-1} = B^{-1}A^{-1} \\ (A^{-1})^{-1} = A \\ (A^T)^{-1} = (A^{-1})^T \end{cases}$$

(3) 梯度:

$$\nabla f(c) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)^T$$

(4) 雅克比矩阵:

$$\begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

多输入多输出函数  $y = f(x)$ , 其中  $y$  与  $x$  分别为  $m$ ,  $n$  维向量。 $y$  的每个分量分别关于  $x$  的每个分量求偏导。

(5) Hessian 矩阵:

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 f}{\partial x_n \partial x_2} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Hessian 矩阵可判断多元函数的凹凸性, 若 Hessian 矩阵正定, 则函数是凸的, 若负定则是凹的。

(6) 正定矩阵:  $x^T A x > 0, x \neq 0$ , 则矩阵  $A$  为正定矩阵。

判断条件:

矩阵的特征值全大于 0

矩阵的所有顺序主子式都大于 0

矩阵合同于单位阵

(7) 二次型:  $x^T A x > 0$

### 1.3 特征值与特征向量

对于  $n$  阶方阵  $A$ , 有  $Ax = \lambda x$ ,  $x$  为特征向量,  $x$  非零,  $\lambda$  为特征值。

(1) 特征值与特征向量的求解:

$$(A - \lambda I)x = 0$$

根据方程组解的理论, 齐次方程组有解, 则  $(A - \lambda I)$  要为不满秩矩阵 (即方程数量小于未知数数量, 方程组有无数个解)。可得  $|A - \lambda I| = 0$ , 解出特征值。

(2) 特征值分解: 正交矩阵  $P, P^{-1} = P^T$ , 使得

$$P^{-1}AP = \Lambda$$

$\Lambda$  为对角矩阵

### 1.4 矩阵和向量的求导

(1)

$$\nabla W^T x = W$$

证明:

根据向量内积公式,  $W^T x = \sum_{i=1}^n w_i x_i$ 。

且对于  $x$  中任意的变量  $x_j$  求偏导:

$$\frac{\partial \sum_{i=1}^n w_i x_i}{\partial x_j} = w_j$$

$$\text{可得: } \nabla W^T x = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = W$$

(2)

$$\nabla x^T A x = (A + A^T)x$$

证明:

$$\nabla x^T A x = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j$$

关于  $x_k$  求偏导, 可能存在两种情况, 即  $i = k, j = k$ , 当  $i = k$  时有:

$$\frac{\partial \sum_{j=1}^n a_{kj} x_k x_j}{\partial x_k} = \sum_{j=1}^n a_{kj} x_j$$

相当于 A 的某一行与 x 的内积，当  $k \in [1, n]$ ，相当于  $Ax$ ；同理当  $j=k$  时，相当于 A 的某一列的转置与 x 内积，即  $A^T x$ 。二者相加，得证。

(3)

$$\nabla^2 x^T A x = A + A^T$$

可由 (1)(2) 式证明。

## 2 概率论

### 2.1 条件概率

$$p(b|a) = \frac{p(a, b)}{p(a)}$$

$$p(a|b) = \frac{p(a, b)}{p(b)}$$

$p(a, b)$  表示 a, b 都发生的概率。 $p(b|a)$  表示 a 发生的情况下，b 发生的概率。

### 2.2 贝叶斯公式

$$p(a|b) = \frac{p(a)p(b|a)}{p(b)}$$

证明：

$$p(a|b)p(b) = p(a)p(b|a) = p(a, b)$$

a 为因，b 为果，因此  $p(b|a)$  称为先验概率， $p(a|b)$  称为后验概率。

### 2.3 随机事件的独立性

$$p(a, b) = p(a)p(b)$$

两个事件独立，表明二者发生的概率没有关联。即：

$$\frac{p(a, b)}{p(a)} = p(b) = p(b|a)$$

## 2.4 期望与方差

### (1) 期望

对于离散随机变量:

$$E(x) = \sum x_i p(x_i)$$

对于连续随机变量 ( $f(x)$  为概率密度函数):

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

### (2) 方差 (反应数据的波动程度)

离散:

$$D(x) = \sum (x_i - E(x))^2 p(x_i)$$

连续:

$$D(x) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x) dx$$

其他计算方式:  $D(x) = E(x^2) - (E(x))^2$

## 2.5 常用分布

### (1) 正态分布 (均值 $\mu$ , 方差 $\sigma^2$ ):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

### (2) 均匀分布:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & x < a, x > b \end{cases}$$

### (3) 二项分布:

$$p(x=1) = p, \quad p(x=0) = 1 - p$$

## 2.6 随机向量

将变量  $x$  扩展成多维向量, 如 2 维  $x = [x_1, x_2]$

$$f(x_1, x_2) \geq 0$$

$$\int_{-\infty}^{\infty} dx_1 dx_2 = 1$$

## 2.7 协方差

协方差反应的是两个随机变量之间线性相关的程度。定义为:

$$\begin{aligned} cov(x_1, x_2) &= E\left((x_1 - E(x_1))(x_2 - E(x_2))\right) \\ &= E(x_1 x_2) - E(x_1)E(x_2) \end{aligned}$$

协方差矩阵 (对为  $n$  维随机向量), 对称的

$$\begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_n) \\ \cdots & \cdots & \cdots & \cdots \\ cov(x_n, x_1) & cov(x_n, x_2) & \cdots & cov(x_n, x_n) \end{pmatrix}$$

## 2.8 最大似然估计

概率分布  $p(x|\theta)$ , 从该分布中采样  $x_i, i = 1, \cdots, l$ , 各个样本之间互相独立。任务: 在已知样本的情况下推测产生这些数据的模型参数  $\theta$ 。因为采样事件已经完成, 我们有理由相信这些样本一定程度上反应了实际的分布, 寻找一个最符合当前观测样本的概率分布。整个采样事件的概率为所有样本概率之积 (似然函数):

$$L(\theta) = \prod_{i=1}^l p(x_i|\theta)$$

最大化整个采样事件的可能性, 但是概率乘积趋近于 0, 且不利于求导, 两边取对数转换成连加函数 (对数似然函数):

$$\ln L(\theta) = \sum_{i=1}^l \ln p(x_i|\theta)$$

$$\max \sum_{i=1}^l \ln p(x_i|\theta)$$

## 3 优化问题

### 3.1 拉格朗日乘子法

有优化问题:

$$\min f(x)$$

$$s.t. \quad h_i(x) = 0, \quad i = 1, \cdots, p$$

将上述带等式约束的优化问题转化为不带约束的优化问题:

$$L(x, \lambda) = f(x) + \sum_{i=1}^p \lambda_i h_i(x)$$

对上式关于  $x$  求导得:

$$\begin{aligned} \nabla_x f + \sum_{i=1}^p \lambda_i \nabla_x h_i &= 0 \\ h_i(x) &= 0 \end{aligned}$$